IN THE CLAIMS:

The following is a complete listing of the claims. This listing replaces all earlier versions and listings of the claims.

Claim 1 (currently amended): An apparatus for processing image data and sound data, comprising:

an image processor operable to process image data recorded by at least one camera showing the movements of a plurality of people to track [[the]] a respective position of each person in three dimensions;

an input unit for inputting voice recognition parameters for each person;

a data store to store data assigning a unique identity to each person tracked by said image processor and to store respective voice recognition parameters for each person;

a sound processor operable to process sound data to determine the direction of arrival of sound;

a speaker identifier operable to determine the unique identity of the person who is speaking by comparing the positions of the people determined by said image processor and the direction of arrival of the sound determined by said sound processor to identify a person at a position in the direction from which the sound arrives;

a voice recognition parameter selector operable to select the voice recognition parameters from said data store of the person identified by said speaker identifier to be the person who is speaking; and

a voice recognition processor operable to process the received sound data to generate text data therefrom using the voice recognition parameters selected by said voice recognition parameter selector.

Claim 2 (canceled)

Claim 3 (previously presented): An apparatus according to claim 1, wherein said image processor is arranged to track each person by processing the image data using camera calibration data defining the position and orientation of each camera from which image data is processed.

Claim 4 (previously presented): An apparatus according to claim 1, wherein said image processor is arranged to track each person by tracking each person's head.

Claim 5 (previously presented): An apparatus according to claim 1, wherein said image processor is arranged to process the image data to determine where at least each person who is speaking is looking.

Claim 6 (previously presented): An apparatus according to claim 1, wherein the received image data comprises a plurality of frames and wherein said speaker identifier is arranged to identify a person who is speaking in a given frame of the received image data using the results of the processing performed by said image processor and said

sound processor for at least one other frame if the speaker cannot be identified using the results of the processing performed by said image processor and said sound processor for the given frame.

Claim 7 (previously presented): An apparatus according to claim 1, further comprising a database for storing at least some of the received image data, the sound data, the text data produced by said voice recognition processor and viewing data defining where at least each person who is speaking is looking, said database being arranged to store the data such that corresponding text data and viewing data are associated with each other and with the corresponding image data and sound data.

Claim 8 (previously presented): An apparatus according to claim 7, further comprising a data compressor for compressing the image data and the sound data for storage in said database.

Claim 9 (previously presented): An apparatus according to claim 8, wherein said data compressor comprises a data encoder for encoding the image data and the sound data as MPEG data.

Claim 10 (previously presented): An apparatus according to claim 7, further comprising a gaze data generator for generating data defining, for a predetermined period, the proportion of time spent by a given person looking at each of the other people during

the predetermined period, and wherein said database is arranged to store the data so that it is associated with the corresponding image data, sound data, text data and viewing data.

Claim 11 (previously presented): An apparatus according to claim 10, wherein the predetermined period comprises a period during which the given person was talking.

Claims 12-16 (canceled)

Claim 17 (currently amended): A method of processing image data and sound data, comprising:

an input step, of inputting voice recognition parameters for each person;

a data storage step, of storing data assigning a unique identity to each person to be tracked in the image data and storing respective voice recognition parameters for each person;

an image processing step, of processing image data recorded by at least one camera showing the movements of a plurality of people to track [[the]] a respective position of each person in three dimensions;

a sound processing step, of processing sound data to determine the direction of arrival of sound;

a speaker identification step, of determining the unique identity of the person who is speaking by comparing the positions of the people determined in said

image processing step and the direction of arrival of the sound determined in said sound processing step to identify a person at a position in the direction from which the sound arrives;

a voice recognition parameter selection step, of selecting voice recognition parameters from among the stored voice recognition parameters of the person identified in the speaker identification step to be the person who is speaking; and

a voice recognition processing step, of processing the received sound data to generate text data therefrom using the voice recognition parameters selected in the voice recognition parameter selection step.

Claim 18 (canceled)

Claim 19 (previously presented): A method according to claim 17, wherein said image processing step includes tracking each person by processing the image data using camera calibration data defining the position and orientation of each camera from which image data is processed.

Claim 20 (previously presented): A method according to claim 17, wherein said image processing step includes tracking each person by tracking the person's head.

Claim 21 (previously presented): A method according to claim 17, wherein said image processing step includes processing the image data to determine where at least each person who is speaking is looking.

Claim 22 (previously presented): A method according to claim 17, wherein the received image data comprises a plurality of frames and wherein said speaker identification step includes identifying a person who is speaking in a given frame of the received image data using the results of the processing performed in said image processing step and said sound processing step for at least one other frame if the speaker cannot be identified using the results of the processing performed in said image processing step and said sound processing step for the given frame.

Claim 23 (previously presented): A method according to claim 17, further comprising a signal generating step, of generating a signal conveying the text data generated in said voice recognition processing step.

Claim 24 (previously presented): A method according to claim 17, further comprising a data storage step, of storing in a database at least some of the received image data, the sound data, the text data produced in said voice recognition processing step and viewing data defining where at least each person who is speaking is looking, the data being stored in the database such that corresponding text data and viewing data are associated with each other and with the corresponding image data and sound data.

Claim 25 (original): A method according to claim 24, wherein the image data and the sound data are stored in the database in compressed form.

Claim 26 (original): A method according to claim 25, wherein the image data and the sound data are stored as MPEG data.

Claim 27 (previously presented): A method according to claim 24, further comprising:

a time proportion data generation step, of generating data defining, for a predetermined period, the proportion of time spent by a given person looking at each of the other people during the predetermined period; and

a time proportion data storage step, of storing the time proportion data in the database so that it is associated with the corresponding image data, sound data, text data and viewing data.

Claim 28 (original): A method according to claim 27, wherein the predetermined period comprises a period during which the given person was talking.

Claim 29 (previously presented): A method according to claim 24, further comprising a generating step, of generating a signal conveying the database with data therein.

Claim 30 (previously presented): A method according to claim 29, further comprising a recording step, of recording the signal either directly or indirectly to generate a recording thereof.

Claims 31-36 (canceled)

Claim 37 (previously presented): A storage device storing computer program instructions for programming a programmable processing apparatus to become configured as an apparatus as set out in claim 1.

Claim 38 (previously presented) A storage device storing computer program instructions for programming a programmable processing apparatus to become operable to perform a method as set out in claim 17.

Claim 39 (previously presented): A signal conveying computer program instructions for programming a programmable processing apparatus to become configured as an apparatus as set out in claim 1.

Claim 40 (previously presented): A signal conveying computer program instructions for programming a programmable processing apparatus to become operable to perform a method as set out in claim 17.

Claim 41 (currently amended): An apparatus for processing image data and sound data, said apparatus comprising:

image processing means for processing image data recorded by at least one camera showing the movements of a plurality of people to track [[the]] a respective position of each person in three dimensions;

input means for inputting voice recognition parameters for each person

storing means for storing data assigning a unique identity to each person tracked by said image processing means and for storing respective voice recognition parameters for each person;

sound processing means for processing sound data to determine the direction of arrival of sound;

speaker identification means for determining the unique identity of the person who is speaking by comparing the position of the people determined by said image processing means and the direction of arrival of the sound determined by said sound processing means to identify a person at a position in the direction from which the sound arrives;

voice recognition parameter selection means for selecting the voice recognition parameters from said storing means of the person identified by said speaker identification means to be the person who is speaking; and

voice recognition processing means for processing the received sound data to generate text data therefrom ~~means~~ using the voice recognition parameters selected by said voice recognition parameter selection means.

Claims 42-92 (canceled)